Machine Learning Basics

Balázs Nagy

2020. február 7.

"The field of study that gives computers the ability to learn without being explicitly programmed."

/Arthur Samuel/

Notes:
h - hypothesis
w - weights
x - input
y - output
ŷ - prediction

m - total number of samples
i - index of samples
C - cost function
MSE - Mean Squared Error
μ - learning rate, $0 < \mu \leq 1$ λ - regularization

$$\underset{3\times 1}{X} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \underset{3\times 1}{Y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \tag{1}$$

$$h_w(x) = wx = \hat{y} \tag{2}$$

$$w = 0, C = 2.33 \tag{3}$$

$$w = 0.5, C = 0.58 \tag{4}$$

$$w = 1, C = 0 \tag{5}$$

$$C(w) = \frac{1}{2m}\sum_{i=1}^{m}(wx^i - y^i)^2 \tag{6}$$

$$w = w - \mu\frac{\partial}{\partial w}C(w) \tag{7}$$

$$\frac{\partial}{\partial w}C(w) = \frac{1}{m}\sum_{i=1}^{m}(wx^i - y^i) \cdot x^i \tag{8}$$

$w_j^t := w_j^{t-1} - \mu\frac{\partial}{\partial w_j}C(w_0, w_1) + \triangle w_j^{t-1}$

$$\begin{bmatrix}(0 \cdot 1 - 1) \cdot 1 \\ (0 \cdot 2 - 2) \cdot 2 \\ (0 \cdot 3 - 3) \cdot 3\end{bmatrix} = \begin{bmatrix}-1 \\ -4 \\ -9\end{bmatrix}, 0.1 \cdot \frac{-14}{3} = -0.46 \tag{9}$$

$$w = 0 - (-0.46) = 0.46 \tag{10}$$

$$\begin{bmatrix}(0.46 \cdot 1 - 1) \cdot 1 \\ (0.46 \cdot 2 - 2) \cdot 2 \\ (0.46 \cdot 3 - 3) \cdot 3\end{bmatrix} = \begin{bmatrix}-0.53 \\ -2.13 \\ -4.8\end{bmatrix}, 0.1 \cdot \frac{-7.46}{3} = -0.249 \tag{11}$$

$$w = 0.46 - (-0.249) = 0.71 \tag{12}$$

$$\begin{bmatrix}(0.71 \cdot 1 - 1) \cdot 1 \\ (0.71 \cdot 2 - 2) \cdot 2 \\ (0.71 \cdot 3 - 3) \cdot 3\end{bmatrix} = \begin{bmatrix}-0.28 \\ -1.13 \\ -2.56\end{bmatrix}, 0.1 \cdot \frac{-3.98}{3} = -0.132 \tag{13}$$

$$w = 0.71 - (-0.132) = 0.842 \tag{14}$$

1. Labor:

Linear regression with one variable

Hypothesis:

$$h_w(x) = w_0 + w_1 x \tag{15}$$

$$h_w(x) = w_0 + w_1 x = \hat{y} \tag{16}$$

Cost function:

$$C = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}^i - y^i)^2 \tag{17}$$

$$C = \frac{1}{2m} \sum_{i=1}^{m} (h_w(x^i) - y^i)^2 \tag{18}$$

$$C(w_0, w_1) \tag{19}$$

$$C(w_0, w_1) = \frac{1}{2m} \sum_{i=1}^{m} (w_0 + w_1 x^i - y^i)^2 \tag{20}$$

$$\underset{m\times 1}{X} = \begin{bmatrix} x^1 \\ x^2 \\ x^3 \\ \vdots \\ x^m \end{bmatrix}, \underset{2\times 1}{W} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}, \underset{m\times 1}{Y} = \begin{bmatrix} y^1 \\ y^2 \\ y^3 \\ \vdots \\ y^m \end{bmatrix} \tag{21}$$

$$X = \begin{bmatrix} 1 & x^1 \\ 1 & x^2 \\ 1 & x^3 \\ \vdots & \vdots \\ 1 & x^m \end{bmatrix} \Rightarrow \underset{m\times 2}{X} = \begin{bmatrix} x_0^1 & x_1^1 \\ x_0^2 & x_1^2 \\ x_0^3 & x_1^3 \\ \vdots & \vdots \\ x_0^m & x_1^m \end{bmatrix} \tag{22}$$

$$\hat{y} = h_w(x) = w_0 + w_1 x^i = w_0 x_0^i + w_1 x_1^i \tag{23}$$

$$\underset{2\times 1}{W} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

$$\underset{m\times 2}{X} = \begin{bmatrix} x_0^1 & x_1^1 \\ x_0^2 & x_1^2 \\ x_0^3 & x_1^3 \\ \vdots & \vdots \\ x_0^m & x_1^m \end{bmatrix} \quad \begin{bmatrix} w_0 x_0^1 + w_1 x_1^1 \\ w_0 x_0^2 + w_1 x_1^2 \\ w_0 x_0^3 + w_1 x_1^3 \\ \vdots \\ w_0 x_0^m + w_1 x_1^m \end{bmatrix} = \underset{m\times 1}{\hat{Y}} = XW$$

$$C = \frac{\sum (XW - Y)^2}{2m} \tag{24}$$

1.1. Gradient descent

To solve: $minC(w_0, \ldots, w_n)$
Algorithm:
repeat until convergence {
$w_j := w_j - \mu \frac{\partial}{\partial w_j} C(w_0, w_1)$
}

Linear Regression Model

$$h_w(x) = w_0 + w_1 x \tag{25}$$

$$C(w_0, w_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_w(x^i) - y^i)^2 \tag{26}$$

Gradient descent
repeat until convergence {
$w_j := w_j - \mu \frac{\partial}{\partial w_j} C(w_0, w_1)$
}

$$\frac{\partial}{\partial w_j} C(w_0, w_1) = \frac{\partial}{\partial w_j} \frac{1}{2m} \sum_{i=1}^{m} (h_w(x^i) - y^i)^2 = \frac{\partial}{\partial w_j} \frac{1}{2m} \sum_{i=1}^{m} (w_0 + w_1 x^i - y^i)^2 \tag{27}$$

$$(j=0) \qquad \frac{\partial}{\partial w_j} C(w_0, w_1) = \frac{1}{m} \sum_{i=1}^{m} (w_0 + w_1 x^i - y^i) \cdot 1 = \frac{1}{m} \sum_{i=1}^{m} (h_w(x^i) - y^i) \cdot x_0^i \tag{28}$$

$$(j=1) \qquad \frac{\partial}{\partial w_j} C(w_0, w_1) = \frac{1}{m} \sum_{i=1}^{m} (w_0 + w_1 x^i - y^i) \cdot x_1^i = \frac{1}{m} \sum_{i=1}^{m} (h_w(x^i) - y^i) \cdot x_1^i \tag{29}$$

Interpretation

$$w_0 = w_0 - \frac{\mu}{m} \sum_{i=1}^{m} (h_w(x^i) - y^i) \tag{30}$$

$$w_1 = w_1 - \frac{\mu}{m} \sum_{i=1}^{m} (h_w(x^i) - y^i) \cdot x^i \tag{31}$$

$$w_0 = w_0 - \frac{\mu}{m} sum(X * w - Y) \tag{32}$$

$$w_1 = w_1 - \frac{\mu}{m} sum(X * w - Y). * X(:, 2) \tag{33}$$

2. Labor:

Linear regression with multiple variable

$$\underset{m\times(n+1)}{X} = \begin{bmatrix} x_0^1 & x_1^1 & x_2^1 & \dots & x_n^1 \\ x_0^2 & x_1^2 & x_2^2 & \dots & x_n^2 \\ x_0^3 & x_1^3 & x_2^3 & \dots & x_n^3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_0^m & x_1^m & x_2^m & \dots & x_n^m \end{bmatrix}, \underset{(n+1)\times 1}{W} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}, \underset{m\times 1}{Y} = \begin{bmatrix} y^1 \\ y^2 \\ y^3 \\ \vdots \\ y^m \end{bmatrix} \tag{34}$$

Hypothesis:

$$h_w(x) = w_0 + w_1x_1 + w_2x_2 + \ldots + w_nx_n \tag{35}$$

$$h_w(x) = w_0x_0 + w_1x_1 + w_2x_2 + \ldots + w_nx_n \tag{36}$$

$$\underset{(n+1)\times 1}{W} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

$$\underset{m\times(n+1)}{X} = \begin{bmatrix} x_0^1 & x_1^1 & x_2^1 & \dots & x_n^1 \\ x_0^2 & x_1^2 & x_2^2 & \dots & x_n^2 \\ x_0^3 & x_1^3 & x_2^3 & \dots & x_n^3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_0^m & x_1^m & x_2^m & \dots & x_n^m \end{bmatrix} \quad \begin{bmatrix} w_0x_0^1 + w_1x_1^1 + w_2x_2^1 + \ldots + w_nx_n^1 \\ w_0x_0^2 + w_1x_1^2 + w_2x_2^2 + \ldots + w_nx_n^2 \\ w_0x_0^3 + w_1x_1^3 + w_2x_2^3 + \ldots + w_nx_n^3 \\ \vdots \\ w_0x_0^m + w_1x_1^m + w_2x_2^m + \ldots + w_nx_n^m \end{bmatrix}$$

Cost function:

$$C(W) = C(w_0, w_1, \ldots, w_n) = \frac{1}{2m}\sum_{i=1}^{m}(h_w(x^i) - y^i)^2 \tag{37}$$

Gradient Descent:

$$w_j := w_j - \mu\frac{1}{2m}\sum_{i=1}^{m}(h_w(x^i) - y^i) \cdot x_j^i \tag{38}$$

$$x = \frac{x - mean(x)}{std(x)} \tag{39}$$

hivatkozás 39

$$v = \begin{bmatrix} v_0 \\ v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

$$v' = \begin{bmatrix} v_0 & v_1 & v_2 & \dots & v_n \end{bmatrix} \quad \begin{bmatrix} (v_0)^2 + (v_1)^2 + (v_2)^2 + \dots + (v_n)^2 \end{bmatrix} \quad \rightarrow v'v = sum(v.\hat{}2)$$

3. Labor:

Logistic regression Linear case

$$y \in \{0, 1\} \tag{40}$$

0: Negative class
1: Positive class

$$h_w(x) = XW \tag{41}$$

Threshold classifier output $h_w(x)$ at 0.5:
If $h_w(x) \geq 0.5$, predict "y=1"
If $h_w(x) < 0.5$, predict "y=0"

$h_w(x)$ can be >1 or <0

Logistic regession: $0 \leq h_w(x) \leq 1$
Logistic Regression Model:
Want $0 \leq h_w(x) \leq 1$
$h_w(x) = g(Xw)$

$$g(z) = \frac{1}{1 + e^{-z}} \tag{42}$$

$h_w(x) \Rightarrow$ estimated probability that $y = 1$ on input x
$h_w(x) = P(y = 1|x, W)$
sigmoid function
Example I.

$$h_w(x) = g(w_0 1 + w_1 x_1 + w_2 x_2) \tag{43}$$

$w = [-3\ 1\ 1]$
Predict: $y = 1$ if $-3 + x_1 + x_2 \geq 0$
$x_1 + x_2 \geq 3$

Example II.

$$h_w(x) = g(w_0 1 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2) \tag{44}$$

$w = [-1\ 0\ 0\ 1\ 1]$
Predict: $y = 1$ if $-1 + x_1^2 + x_2^2 \geq 0$
$x_1^2 + x_2^2 \geq 1$

$g(z) \geq 0.5$
when $z \geq 0$

$$Cost(h_w(x), y) = \begin{Bmatrix} -log(h_w(x)),\ if\ y = 1 \\ -log(1 - h_w(x)),\ if\ y = 0 \end{Bmatrix} \tag{45}$$

$$Cost(h_w(x), y) = -y \cdot log(h_w(x)) - (1 - y) \cdot log(1 - h_w(x)) \tag{46}$$

$$C(W) = -\frac{1}{m} \sum_{i=1}^{m} y^i \cdot log(h_w(x^i)) + 2(1 - y^i) \cdot log(1 - h_w(x^i)) \tag{47}$$

Want min $min_W\{C(W)\}$
Algorithm:
repeat until convergence {
$W_j := W_j - \mu \frac{\partial}{\partial W_j} C(W)$
}

$$\frac{\partial}{\partial W_j} C(W) = \frac{1}{m} \sum_{i=1}^{m} (h_w(x^i) - y^i) \cdot x_j^i \tag{48}$$

4. Labor:

Logistic regression Non Linear case

Using Polinomial Features
$x_1\ x_2 \ \Rightarrow\ 1\ x_1\ x_2\ x_1^2\ x_1x_2\ x_2^2\ x_1^3\ x_1^2x_2\ x_1x_2^2\ x_2^3$

$$w_0 + w_1 x$$

"Underfit"
"High Bias"

$$w_0 + w_1 x + w_2 x^2$$

Just Right

$$w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4$$

"Overfit"
"High Variance"

Regularization:

$$C(w) = \frac{1}{2m}\sum_{i=1}^{m}(h_w(x^i) - y^i)^2 + \lambda \sum_{j=1}^{n} w_j^2 \tag{49}$$

If λ large: algorithm result in underfitting
(fails to fit even the training set)

Regularized Logistic Regression:
Repeat{

$$w_0 := w_0 - \mu \frac{1}{m}\sum_{i=1}^{m}(h_w(x^i) - y^i)\cdot x_0^i$$

$$w_j := w_j - \mu[\frac{1}{m}\sum_{i=1}^{m}(h_w(x^i) - y^i)\cdot x_j^i] + \frac{\lambda}{m} w_j$$

}

Cost function and derivative:

$$C(w) = [-\frac{1}{m}\sum_{i=1}^{m} y^i \cdot log(h_w(x^i)) + (1-y^i) \cdot log(1-h_w(x^i))] + \frac{\lambda}{2m}\sum_{j=1}^{n} w_j^2 \quad (50)$$

$$\frac{\partial}{\partial w_0} C(w) = \frac{1}{m}\sum_{i=1}^{m}(h_w(x^i) - y^i) \cdot x_0^i + 0 \quad (51)$$

$$\frac{\partial}{\partial w_j} C(w) = \frac{1}{m}\sum_{i=1}^{m}(h_w(x^i) - y^i) \cdot x_j^i + \frac{\lambda}{m} w_j \quad (52)$$

5. Labor:

Multi Class Classification

6. Labor:

Neural Networks Basics

7. Labor:

Neural Network Train

To the picture

$$x_1^{(1)}$$
$$x_2^{(1)}$$
$$x^{(1)}$$
$$w_{01}^{(1)}$$
$$w_{02}^{(1)}$$
$$w_{03}^{(1)}$$
$$w_{11}^{(1)}$$
$$w_{12}^{(1)}$$
$$w_{13}^{(1)}$$
$$w_{21}^{(1)}$$
$$w_{22}^{(1)}$$
$$w_{23}^{(1)}$$
$$w_{01}^{(2)}$$
$$w_{02}^{(2)}$$
$$w_{03}^{(2)}$$
$$w_{11}^{(2)}$$
$$w_{12}^{(2)}$$
$$w_{13}^{(2)}$$
$$w^{(1)}$$
$$w^{(2)}$$
$$x^{(1)}$$
$$s^{(2)}$$

$$s^{(3)}$$

$$a^{(2)}$$

$$a^{(3)}$$

$$s = \sum_{i=1}^{n} w_i \cdot x_i \tag{53}$$

$$y = a(s) \tag{54}$$

$$xw^{(1)} = s^{(2)} \tag{55}$$

$$a^{(2)} = f(s^{(2)}) = sigmoid(s^{(2)}) \tag{56}$$

$$s^{(3)} = a^{(2)}w^{(2)} \tag{57}$$

$$\hat{y} = f(s^{(3)}) = sigmoid(s^{(3)}) \tag{58}$$

$$C = \sum\{\frac{1}{2}(y - \hat{y})^2\} \tag{59}$$

$$C = \sum\{\frac{1}{2}(y - a^{(3)})^2\} \tag{60}$$

$$C = \sum\{\frac{1}{2}(y - f(s^{(3)}))^2\} \tag{61}$$

$$C = \sum\{\frac{1}{2}(y - f(a^{(2)}w^{(2)}))^2\} \tag{62}$$

$$C = \sum\{\frac{1}{2}(y - f(f(s^{(2)})w^{(2)}))^2\} \tag{63}$$

$$C = \sum\{\frac{1}{2}(y - f(f(xw^{(1)})w^{(2)}))^2\} \tag{64}$$

Back propagation:

$$\frac{\partial C}{\partial w^{(2)}} = \frac{\partial \sum \frac{1}{2}(y - \hat{y})^2}{\partial w^{(2)}} = \sum(\frac{\partial \frac{1}{2}(y - \hat{y})^2}{\partial w^{(2)}}) \tag{65}$$

$$\begin{aligned} \frac{\partial \frac{1}{2}(y - \hat{y})^2}{\partial w^{(2)}} &= (y - \hat{y})(-\frac{\hat{y}}{\partial w^{(2)}}) \\ &= -(y - \hat{y}) \cdot \frac{\partial \hat{y}}{\partial s^{(3)}} \cdot \frac{\partial s^{(3)}}{\partial w^{(2)}} \\ &= -(y - \hat{y}) \cdot f'(s^{(3)}) \cdot \frac{\partial a^{(2)}w^{(2)}}{\partial w^{(2)}} \\ &= \delta^{(3)} \cdot a^{(2)} \end{aligned} \tag{66}$$

Dimension check:

$$(a^{(2)})^T \delta^{(3)} \tag{67}$$

$$-(y - \hat{y}) \cdot f'(s^{(3)}) = \delta^{(3)} \tag{68}$$

$$(a^{(2)})^T \delta^{(3)} = \frac{\partial C}{\partial w^{(2)}} \tag{69}$$

$$\delta^{(3)} \cdot (w^{(2)})^T \cdot f'(s^{(2)}) = \delta^{(2)} \tag{70}$$

$$x^T \delta^{(2)} = \frac{\partial C}{\partial w^{(1)}} \tag{71}$$

$$w^{(1)} = w^{(1)} - \mu \frac{\partial C}{w^{(1)}} + regularization \tag{72}$$

$$w^{(2)} = w^{(2)} - \mu \frac{\partial C}{w^{(2)}} + regularization \tag{73}$$

$\delta_1^{(3)}$

$\delta_1^{(2)}$

$\delta_2^{(2)}$

$\delta_3^{(2)}$

8. Labor:

Regularization

9. Labor:

Support Vector Machine

Logistic regression:

$$h_w(x) = \frac{1}{1+e^{-Xw}} \tag{74}$$

$$h_w(x) = g(Xw) \tag{75}$$

$$h_w(x) = g(z) \tag{76}$$

Cost function:

$$C = -(y \cdot log(h_w(x) + (1-y) \cdot log(1-h_w(x))) \tag{77}$$

$$C = -y \cdot log(h_w(x) - (1-y) \cdot log(1-h_w(x)) \tag{78}$$

If $y = 1$, we want $Xw \geq 1$ (not just ≥ 0)
If $y = 0$, we want $Xw < -1$ (not just < 0)

$$h_w(x) = \frac{1}{1+e^{-Xw}} = \frac{1}{1+e^{-z}} = g(z) \tag{79}$$

$$cost_1(z)$$

$$cost_0(z)$$

$$min_w \, \frac{1}{m}\Big[\sum_{i=1}^{m} y^{(i)} \cdot \big(-log(h_w(x^{(i)}) + (1-y^{(i)})\big) \cdot \big(-log(1-h_w(x^{(i)}))\big)\Big] + \frac{\lambda}{2m}\sum_{j=1}^{n} w_j^2 \tag{80}$$

$$min_w \, C\Big[\sum_{i=1}^{m} y^{(i)} \cdot cost_1(h_w(x^{(i)}) + (1-y^{(i)})\big) \cdot cost_0(1-h_w(x^{(i)}))\big)\Big] + \frac{1}{2}\sum_{j=1}^{n} w_j^2 \tag{81}$$

10. Labor:
Spam Email

11. Labor:

K-Means

12. Labor:

Principal Component Analysis

13. Labor:

Anomaly Detection

14. Labor:

Recommender System